International Journal of Computer Networking,
Wireless and Mobile Communications (IJCNWMC)
ISSN(P): 2250-1568; ISSN(E): 2278-9448
Vol. 5, Issue 6, Dec 2015, 19-26

TRANS STELLAR
•Journal Publications •Research Consultancy

A STUDY ON BIG DATA MODELING TECHNIQUES

B. SAI JYOTHI & S. JYOTHI

[1]*Research Scholar, Sri Padmavathi Mahila University, Tirupathi, Andhra Pradesh, India*

[2]*Professor, Sri Padmavathi Mahila University, Tirupathi, Andhra Pradesh, India*

ABSTRACT

Big data describe a gigantic volume of both structured and unstructured data. Big data may generated from sensors and social networking web sites such as Facebook and Twitter. Big Data is popularly known to deal with the data with the 5 characteristics, Volume, Velocity, Variety, Veracity and Value. Conventional systems cannot be used efficiently for storing such vast amount of data exhibiting these 5V's. In this scenario, there is a need for a system that can manage such data.

In literature, most of the times, modeling is considered as a substitute for "entity relationship modeling." But modeling is not only limited to relational databases, it can also be efficiently utilized to communicate ideas beyond this. Data modeling is also used to design data structures at various levels of abstraction from conceptual to physical. Modeling often is used to describe logical design of the system, where as design describes physical implementation of a database. Therefore, data modeling will also lead to effective design.

Most of the organizations are interested in gathering, storing and analyzing this data because it can add significant value to the decision making process. Data Modeling plays a crucial role in big data analytics because 85% of big data is unstructured data. Hence it should modeled as required to the organization needs. And important component of a Big Data application is the data model in which the Big Data resides. And the big data model should provide a visual way to manage data resources, and creates fundamental data architecture so that it can have more applications to optimize data reuse and reduce computing costs. This paper is an attempt to analyse the various big data modeling techniques.

KEYWORDS: Social Networks, Big Data, Data Model

Received: Oct 15, 2015; **Accepted:** Oct 22, 2015; **Published:** Nov 19, 2015; **Paper Id.:** IJCNWMCDEC20153

Original Article

INTRODUCTION

Big Data deals with the data that have high Volume, high Velocity, high Variety, high Veracity and Value also known as 5V's of Big Data. *Volume* depicts very large amount of data generated every fraction of second. With this amount, it becomes difficult to store and analyze many of the datasets. Today's world deal mostly with heterogeneous data, which originates from many different sources. This introduces *Variety* in data. Measures were taken to make most unstructured data as structured which lead to evolution of relational data. These measures have fallen short, as 80% of the world's data is unstructured and comes in streams(text, images, video, voice, etc.). This creates necessity of new technology and tools named big data technology which brings together different types of data. Not only variety, data is also being generated with enormous speed, which is the third V of Bigdata, *Velocity.* According to a recent survey, just the facebook generates 500 TB of data per day. As the data made available on internet is by different people, there is a high possibility of having biases and noise in

data, which is more challenging compared to the other characteristics. This characteristic is known as *Veracity*. The existing data analysis tools cannot capture volume, velocity, variety and veracity of data. Moreover, there is hidden knowledge in the raw data, which is valuable in organizational decision making, defines the fifth V, *Value*.

Recently, Big Data has been getting increasingly attention and recognition due to its broad research and application prospects. As the interest in big data has risen rapidly, there is an increasing effort in trying to analyze and store it. Earlier most of the DBMS packages manages structured data only but data have some other unstructured component also and during analyses that unstructured data component plays a very important role in decision making for example web logs makes a very large component of website data and is generally unstructured but in order to understand the interest of the customer it is an important resource. Today the majority of data originated is unstructured and managing that data is still a challenge.

Hence, before applying the analytics on big data , it should be modeled. Modeling the big data is important because it contains structured, semi structured, unstructured data. And 85% of data is unstructured and semi-structured. To map the all these varieties of data, modeling plays an important role in big data analytics.

DATA MODELING TECHNIQUES

Data Modeling in Social Networks: The model proposed in [1] is the data model based on big table. Figure 1 shows the data model that uses Big Table by Google to store social network data such as comments and contents. This table can be viewed as a Key/value based model. It has *n* rows and each row has a unique identifier in the Row key field. Each Row key has several columns. In each column, column-key and column-value is stored. In a column, *n* key-value pairs exist, and the column-key is used to identify each data unique.

Row key	General Info	Column families(Comment List)			
Content_ID	Title: Data&time: Content: Type:	Comment_ID	Comment_ID	Comment_ID	...
		User_ID: Date&time: Content: Rating:	User_ID: Date&time: Content: Rating:	User_ID: Date&time: Content: Rating:	...

Figure 1: Big Table Model

Figure 2 shows an example of the data model. In the table, the data model is based on key/value and comprised of Keyspace, Column Family, Column, or Value. The Keyspace simply refers to a tag to put together Column Family, and Column Family is a structure having Columns. This is the different from RDBMS, because data can be stored in a column in a form of list or key/value.

Row key	General Info	Column families			
		Comment_01	Comment_02	Comment_03	...
Con_101	Title: Data & time: 20120602- 12:24:35 Content: Android Type:	User_ID:101 Date & time: 20120602- 15:50:32 Content: Android Rating:5	User_ID:102 Date & time: 20120602-16:21:32 Content: Android Rating:4	User_ID:104 Date & time: 20120602- 16:21:50 Content: Android Rating:2	...

Figure 2: Example Data in the Content Table

This proposed model can be used in information recommendation in Social Networks.

Data Modeling in Cloud Environment: The model proposed in [2] models the data in cloud environment. In this model first a schema for Big Data is build. For creating schema this method first recognizes type of the data which is arriving from multiple sources. If the identified data is unstructured then the key information from it is obtained by developing metadata. As to develop metadata it extracts the entities including information about names, publisher etc and extracts the facts including the information about the type of content, issues etc. Metadata Development is based on Dublin Core Metadata development [3]. Using Dublin Core Metadata development 15 element were used to develop metadata for each of the unstructured data. Figure 3 depicts those 15 elements.

1. Title	2. Explanation	3. Date	4. Identifier
5. Creator	6. Publisher	7. Type	8. Source
9. Subject	10. Contributor	11. Format	12. Language
13. Relation	14. Location	15. Rights	

Figure 3: 15 elements in DCME

Once the required information is extracted it will be categorized according to the type of data and the table is created that will be mapped with the structured data schema. After mapping both the schemas a unified schema is prepared, which will hold data about all the data stored in the database in Big Data Dictionary.

At the Cloud Storage level data is stored on commodity hardware by using the Hadoop's HDFS. The clusters are formed on the basis of the type of data , two types of clusters will be formed on to store the structured data and the other to store the unstructured data. Among the cluster of unstructured data further clustering will be done on the basis of category of unstructured data for example if three category of unstructured data like text, audio and video then three clusters will be formed for that.

Ontology Based Big Data Model: The model proposed in [4] is a model based on Ontology. An ontology is an explicit specification of a conceptualization [5], are abstract modeling of the things in the real world such as concepts, constraints, and identity. Due to the big data integrated from heterogeneous data sources, the data could not be shared and understand each other. The following analysis is how to combine ontology technology, to build a data model conforms in line with the MapReduce framework, to solve problems of unstructured data.

The method of constructing big data model based on ontology: OWL (Web Ontology Language) which is a W3C recommended standard ontology description language is used to develop a data model. Using OWL *First* construct class of the model. For example

<owl:class rdf:about="product" />

<owl:class rdf:about="deal" />

<owl:class rdf:about="participant" />

And *second*, construct structural properties of the model. For example

<owl:class rdf:about="computer" >

<rdfs:subClassOf rdf:resource="product" />

</owl: class>

<owl:class rdf:about="notebook" >

<rdfs:subClassOf rdf:resource="computer" />

</owl: class>

And *finally* construct individuals of the model. For example

<owl:NamedIndividual rdf:about="T430">

<rdf: type rdf:resource="notebook" />

<hasCPU> i5-3210M </ hasCPU>

<hasRAM> 8G </ hasRAM>

<hasHD> 500G </ hasHD>

<hasPIC> T430/pic001.jpg </ hasPIC>

<hasVID> T430/vid001.avi </ hasVID>

<hasCOM> T430/com001.txt </hasCOM>

</owl: NamedIndividual>

Ontology-based Key/Value Storage Model:

Big data's storage, query, analyze system should have the features of high scalability (to meet the growing needs to the amount of data), high performance (to meet the real-time and high-performance query processing data for reading and writing), fault tolerance (to ensure the availability of distributed systems), scalability (distribution according to need resources) In recent years, a new model for big data management, resulting in Google's BigTable , based on Hadoop HDFS of HBase [5] provides a platform for technical support addressing the big data storage, query and analysis. The data is stored in a Key / Value model, could be handled directly by HBase NoSQL databases which is easy for data updating dynamically, and meet the needs of high concurrent data processing in a big data environment. The following tables shows an instance of HBase tables.

Table 1: Instance of Data Storage in Key/Value Model

Key	Value					
	hasCPU	hasRAM	hasHD	hasPIC	hasVID	hasCOM
URI#T430-1/	i5-3210M	8G	500G	URI#T430/001.jpg	URI#T430/001.avi	URI#T430/001.txt
URI#T430-2/	i5-3210M	8G	500G	URI#T430/002.jpg	URI#T430/002.avi	URI#T430/002.txt
URI#T430-3/	i5-3210M	8G	500G	URI#T430/003.jpg	URI#T430/003.avi	URI#T430/003.txt

Table 2: HBClass Storage Structure

Row-Key Class	Column Family SubClass	Column Family Property
Product	Phone	Phone_Parameter
	Camera	Camera_Paramete
	Computer	Computer_Parameter
Deal	Manufacturer	hasProduce
	Saler	hasSale
	Buyer	hasBuy
Participant	PIC	hasPIC
	VID	hasVID
	COM	hasCOM

Table 3: HBProperty Storage Structure

Row-Key Parameter	Column Family SubProperty	Column Family Individual
Phone_Parameter	hasCPU	CPU GHZ
	hasRAM	RAM G
	HasROM	ROM M
Camera_Parameter	hasPixel	Pixel X
	hasVR	VR N
	hasCMOS	CMOS M*N
Computer_Parameter	hasCPU	CPU GHZ
	hasRAM	RAM G
	hasHD	HD G

Table 4: HBInstance Instance Table

Row-Key Instance	Column Family HBClass/ Product	Column Family HBProperty	Column Family Deal/ Manufacturer	Column Family Participant/ COM
URI#/Phone/Mi2/	Phone	Phone_Parameter	China	This is a good phone.
URI#/Camera/D80/	Camera	Camera_Paramete	Japan	D80 is Stop production.
URI#/Computer/T430/	Computer	Computer_Parameter	China	It's my best friend!
URI#/Office/hp2050/	Office	Office_Parameter	USA	My office use it.

NLP and Ontology Modeling for Unstructured Data

The model proposed in [6] models the unstructured data using NLP with Ontology from multiple sources is unstructured. unstructured data means, it doesn't have specific structure but it has grammar. unstructured data includes user feedbacks in commercial websites, comments, advertisements, graphics, text, emails etc. By using natural language processing text on a web page can expressed in specific structure. A language is a system of communication which has grammatical rules to express thoughts. Based on the grammar the NLP extracts the information in unstructured data and store it in organization specific format. NLP techniques like Anchoring, Future pacing, Swish, Reframing, Ecology are used to extract the meaning in unstructured data and store it in a structured format. Text in NLP is handled in different layers which are shown in the Figure 4.

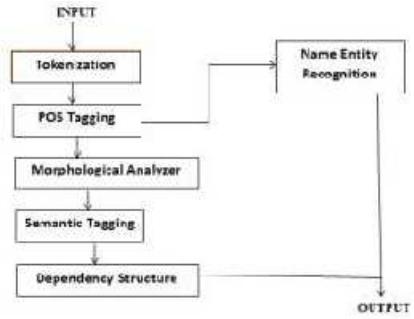

Figure 4: NLP Layers

Tokenization defines the individual words involved in the sentence or paragraph. Part of Speech Tagging describes the category and role of a word. Morphological Analyser shows the root word along with its features such as gender, case etc. Semantic Tagging refers to assigning sense to a word or phrase and can be connected with Semantic Web for semantic annotation and for the ontologies development also. Dependency phrase defines the role of the object in relation to others. NLP attempts to enable computers to make sense of human language. System machinery works on patterns for processing. Intelligent agents make use of NLP layers for fetching patterns from data. The next section includes how layers support for unstructured data.

NLP and ontologies can be often seen contrary to each other. As above, ontology model is a classification of entities and models the relationship among those entitles while the aim of NLP is to identify the entities and understanding of relationship among those entities. Using ontologies with NLP, understanding of natural language through systems become smarter enough to make inference and respond with defined and relevant result what a user requests. When, working with domain ontology, collection and description of concepts collected for domain corpus where ontology concepts exist. The process is repeated for extending the ontology with relevant related concepts. Then a relationship between concepts is established for showing concepts dependency. Concepts relationship process is an important phase for expanding the domain store and as well as for covering at most related terminologies in the specific domain. Dependency structure, one of the NLP techniques can be applied for showing specific ontology domain concepts dependency. For example, there is a sentence which is related to computer science domain:“on-screen keyboard displays a virtual keyboard on computer-screen.” The dependency structure for the sentence is shown in figure 5.

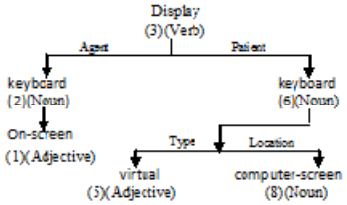

Figure 5: Dependency Structure

A dependency relation is represented by a tree structure as shown above. As figure 3 shows: “display” is the head both “keyboards” and dependents are shown at the lower ends of arrows. The numbers in the tree represent location of each word in the sentence. Using these dependencies, connections between ontology concepts can be evaluated. These are:

1: nadj (2-keyboard, 1-on-screen)

2: nsubj (3-display, 2- keyboard)

3: nobj (3-display, 6- keyboard)

4: nadj (6- keyboard, 5-virtual)

5:prep_on(6-keyboard, 8-computer-screen)

Appropriate concepts that have relevance to the domain ontology are extracted from these connectives. Here, extracted ontology concepts are:

1-on-screen keyboard

2-virtual keyboard and

3-computer-screen

The above example is for the modelling of domain ontology. Likewise, Ontology acquisition process can be applied to make ontology store huge.

CONCLUSIONS

The paper presented above gives an understanding of Big Data modeling techniques, along with it the paper gives a review of the research and developments in the field of Big Data and Modeling techniques. The paper also provides the suggestion regarding the future researches in the field of Big Data and Modeling..

REFERENCES

1. *Xiaoyue Han, Lianhua Tian, Minjoo Yoon, Minsoo Lee, "A Big Data Model supporting Information Recommendation in Social Networks", 2012 Second International Conference on Cloud and Green Computing. IEEE.*
2. *Imran Khan, S. K. Naqvi, Mansaf Alam, S. N. A Rizvi, "Data Model for Big Data in Cloud Environment", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE.*
3. *Abdullah, M. F., & Ahmad, K. (2013, November). The mapping process of unstructured data to structured data. In Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on (pp. 151-155). IEEE.*
4. *Li Kang, Li Yi, LIU Dong, "Research on Construction Methods of Big Data Semantic Model", Proceedings of the World Congress on Engineering 2014 Vol- I, WCE 2014, July 2 - 4, 2014, London, U.K. IEEE.*
5. *Gruber TR, "A translation approach to portable ontology specifications[J]" Knowledge acquisition, 1993, 5(2): 199-220.*
6. *Hemant Kumud, "Handling Unstructured Data for Semantic Web – A Natural Language Processing Approach", Scholars Journal of Engineering and Technology (SJET) ISSN 2321-435X (Online) Sch. J. Eng. Tech., 2014; 2(2A):193-196*